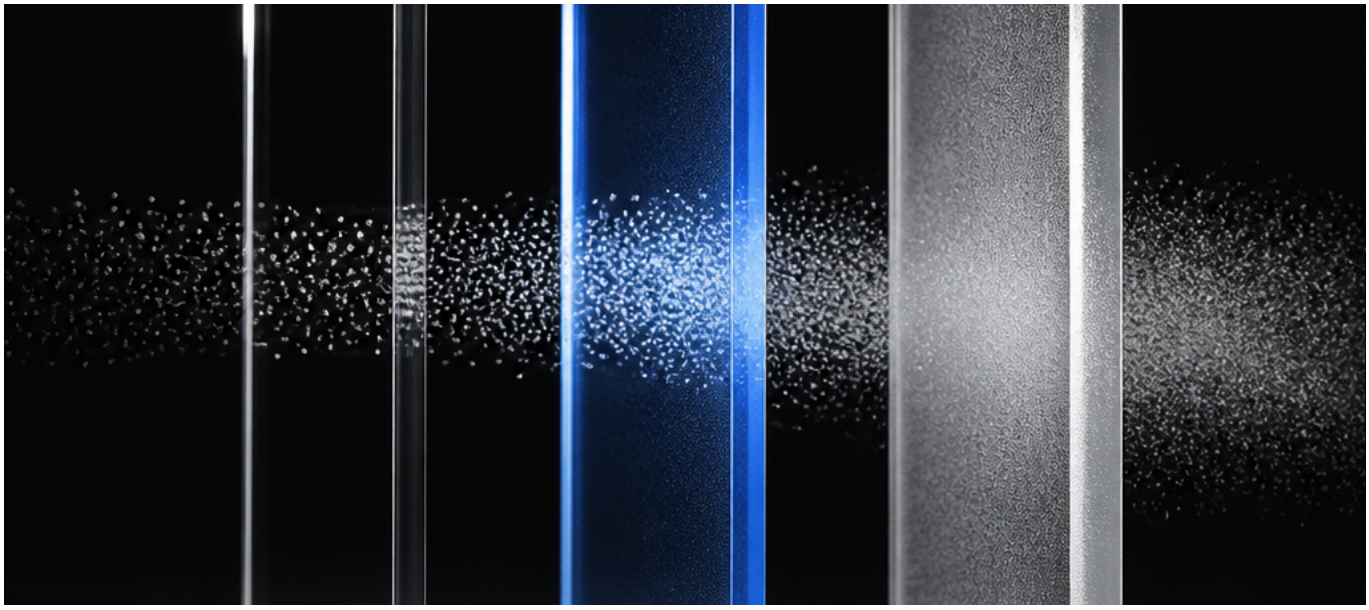


Bedrock Guardrails vs Azure Prompt Shields vs LlamaFirewall: The Honest Runtime AI Firewall Comparison for CTOs

$0.08 versus $0.375. Same job, prompt-attack detection on inbound text, and a 4.7x price gap between AWS and Azure. Neither vendor publishes a detection rate you can hold up against the other.

Anyone picking a runtime AI firewall in late 2026 is choosing without the one number that should decide it. AWS prices its prompt-attack safeguard at $0.08 per 1,000 text units under classic text-only pricing; Azure Prompt Shields lists at roughly $0.375 to $0.38 per 1,000 text records, with 5,000 free each month. That 4.7x gap is the decisive figure, because no comparable detection rate exists for any of the three.

AWS updated its Bedrock Guardrails documentation on 6 November 2026 with malicious-code-injection detection and PII protection inside code structures. Meta refreshed the LlamaFirewall docs on 3 October 2026. Artificial Analysis published a guardrail safety

benchmark updated 11 November 2026. None of that produces a side-by-side detection comparison, because no two of these three are measured on the same datasets under the same conditions. Price and deployment model are what you actually get to decide on.

What the three vendors publish, attribute by attribute

ATTRIBUTE	BEDROCK GUARDRAILS	AZURE PROMPT SHIELDS	LLAMAFIREWALL
Prompt-attack price	**$0.08 / 1,000 text units (classic text-only)**	~$0.375–$0.38 / 1,000 text records	**$0 vendor fee (self-hosted)**
Other filters	$0.15 / 1,000 units (content filters, denied topics); $0.10 (sensitive info, contextual grounding); word and regex filters free	n/a (two attack classes only)	n/a (no per-request pricing)
Free allowance	None	**5,000 text records / month**	**No metering**
Billing unit	1 text unit = up to 1,000 characters; each enabled safeguard bills separately	Per text record; 10,000 characters per prompt, max 5 documents	**Hosting and inference hardware**
Attack coverage as documented	Prompt attacks, malicious code injection, PII in code structures (6 Nov 2026)	User-prompt attacks, document attacks	**Direct and indirect prompt injection, including obfuscated and tool-response attacks (3 Oct 2026)**
Published latency	n/a	n/a	~350ms P50, third-party, hosted Llama Guard 4 endpoint only
Comparable detection rate	n/a	n/a	n/a

The bottom row is the honest one. Three products, one shared job, and the column that should drive the decision is empty across the board.

The benchmarks that exist answer different questions

Artificial Analysis uses average F1 across three datasets as its headline metric and measures latency end-to-end, one request at a time, discarding 20 warm-up requests. That is a defensible methodology, and a specific one. F1 averaged over a particular dataset mix rewards a particular balance of recall and precision, so a vendor tuned for a different threshold will look worse without being worse for your traffic.

Separately, guardrailprobe fires 78 adversarial probes across all ten OWASP LLM Top-10 categories, including 21 for insecure output handling and 12 for sensitive information disclosure. Different question, different answer shape. A score on guardrailprobe and a score on the Artificial Analysis benchmark are not two readings of the same thing.

And the number everyone quotes in procurement decks, 95% detection at 1% false positives, is a derived lab illustration from an input-guardrails writeup. It is not a published result for Azure, AWS or Meta. If you have seen it in a vendor comparison, someone laundered an example into a claim.

The 4.7x gap is narrower than it looks, and also wider

Narrower, because Azure gives you 5,000 text records a month free and AWS gives you nothing. At low volume Azure can be cheaper in absolute terms. Wider, because of how AWS counts. One text unit is up to 1,000 characters, so a 5,600-character input bills as six units, and each enabled safeguard bills separately. Turn on content filters ($0.15), denied topics (same meter), sensitive-information checks ($0.10) and the prompt-attack filter ($0.08), and your per-request cost becomes the sum of every meter you lit up, multiplied by six for that one long prompt.

Azure's limits cut the other way: 10,000 characters per prompt and at most five documents. If your RAG context routinely exceeds that, you are chunking before the shield sees it, which changes what the shield is actually inspecting.

MY READ

The price comparison only holds if you model your real traffic shape: average prompt length,

number of safeguards enabled, and document count per request. I would build that spreadsheet before reading another feature matrix. On AWS the free word filters and regex-based sensitive-information filters are worth more than they look, because ***deterministic rules you do not pay per call for are the cheapest layer you will ever run****.*

LlamaFirewall's zero vendor fee is not free

No per-request vendor fee means the cost moves entirely to hosting and inference hardware. That is a real number, just one you have to compute yourself. The single latency figure in circulation, roughly 0.35s P50, is third-party and measured on a hosted Llama Guard 4 endpoint rather than the full LlamaFirewall stack. Treat it as a floor, not a spec.

What LlamaFirewall documents and the others do not is indirect prompt injection via tool responses, including obfuscated and tool-response attacks. If your agents call tools and feed the output back into the model, that is the attack surface I would weigh hardest. I have written before about Apache-2.0 prompt-injection filtering at the tool-call boundary, and the same logic applies: the dangerous text usually arrives from a system you trust rather than from the user.

Pick by where your model already lives

If you are already on Bedrock and running a handful of safeguards on short prompts, Bedrock Guardrails is the cheapest path, and the 6 November addition of malicious-code-injection detection closes a gap that mattered if you generate or execute code. Model your character counts first.

If you are on Azure, Prompt Shields covers user-prompt attacks and document attacks, nothing else, so budget for a second layer. The 5,000 free records a month make it painless for a pilot, and the list price makes it expensive at scale.

If you run agents with tool calls, or you need the firewall outside any single cloud, LlamaFirewall is the only one of the three whose documentation names obfuscated and tool-response injection. You pay for it in operations rather than per request.

Run your own 78 probes before you sign anything

My take: run guardrailprobe's 78 probes against each candidate yourself, on your own traffic patterns. It is the only way I know to get numbers that are comparable, because you control the dataset and the thresholds. The vendors will not hand you that, and the published benchmarks answer adjacent questions. I am not certain a self-run probe suite settles it either, since 78 probes is a sample and your real attackers are not. For a reference point on what a standardized probe suite does and does not buy you, my piece on garak 0.17.1 and NVIDIA's EU AI Act mapping covers the same trap from the compliance side.

Then decide on price and deployment, because as of November 2026 those are the only criteria with defensible numbers behind them.

WHAT TO DO WITH THIS

Do not let a vendor sell you on a detection rate none of the three publish comparably. Compute your cost from real character counts and enabled safeguards, run your own probe suite, and choose on deployment model. If you want a second pair of eyes on that cost model before you commit: Book a call →



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →